

本周周报（2013. 5. 6-2013. 5. 12）

郭方舟

本周工作

1. 继续进行 vast challenge 3 的数据调研，对数据的分布进行了统计。

目前使用了第一周的 bbexport 数据和网络流数据三个文件中的一个文件。下面先介绍统计结果。

1) 数据量

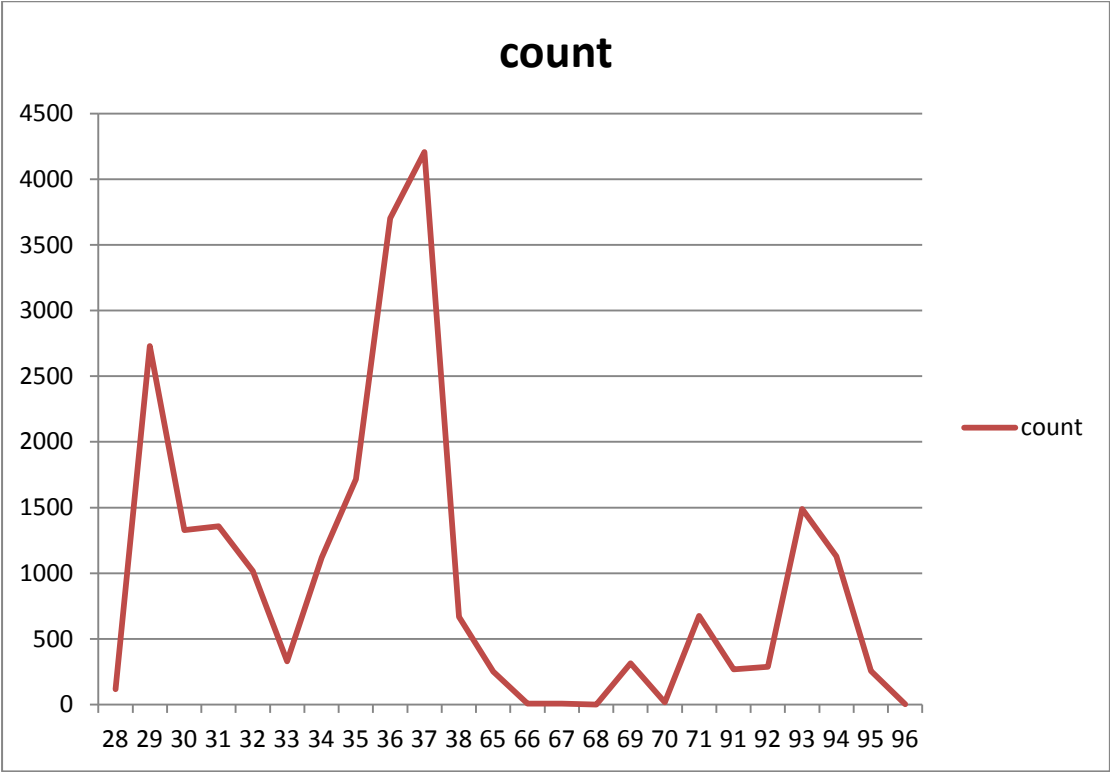
文件名	bbexport.csv	nf1.csv
数据量（条）	340 万	1100 万

2) bbexport.csv 中数值字段的分布情况，-1 意为没有数据。

a. diskUsagePercent

diskUsagePercent	count
-1	2717085
27	667888
28	117
29	2729
30	1328
31	1357
32	1017
33	329
34	1118
35	1716
36	3701
37	4207
38	669
65	251
66	7
67	8
68	1
69	315
70	16
71	676
91	269
92	287
93	1490
94	1128
95	256
96	2

这一字段的分布为：



b. pageFileUsagePercent

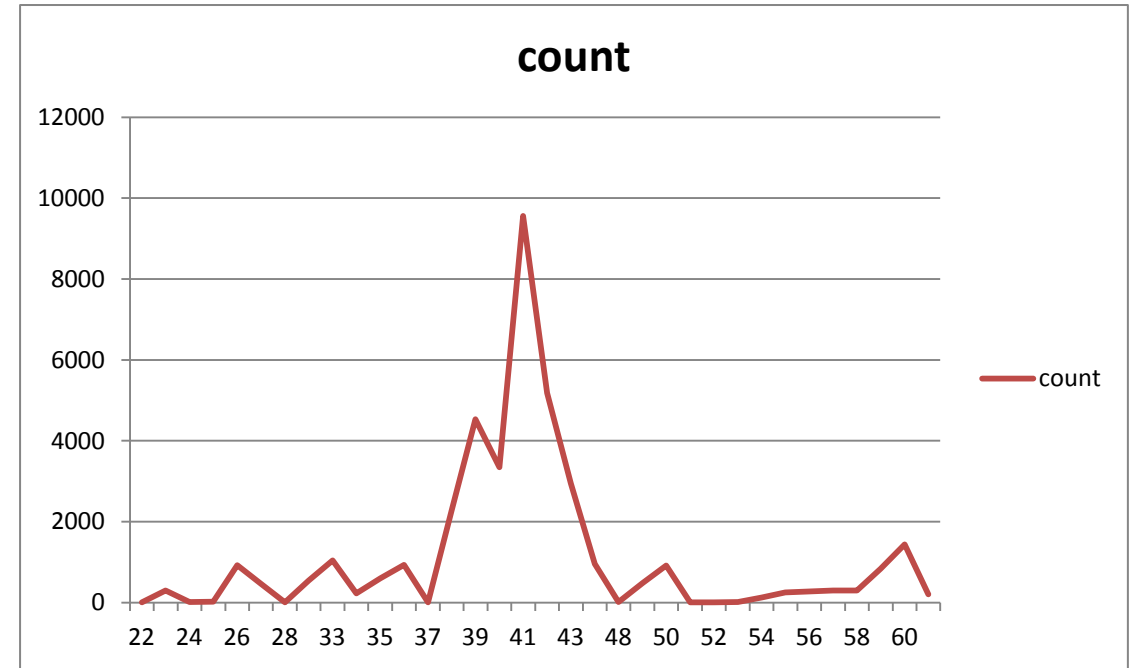
-1	3368968
0	37285
1	346
2	1368

c. numProcs

numProcs	count
-1	3368964
22	3
23	296
24	11
25	20
26	921
27	461
28	2
32	540
33	1045
34	222
35	596
36	929
37	6
38	2286
39	4533

40	3343
41	9562
42	5180
43	2949
44	956
48	12
49	482
50	912
51	3
52	2
53	9
54	120
55	249
56	275
57	299
58	299
59	839
60	1435
61	206

其分布如下：

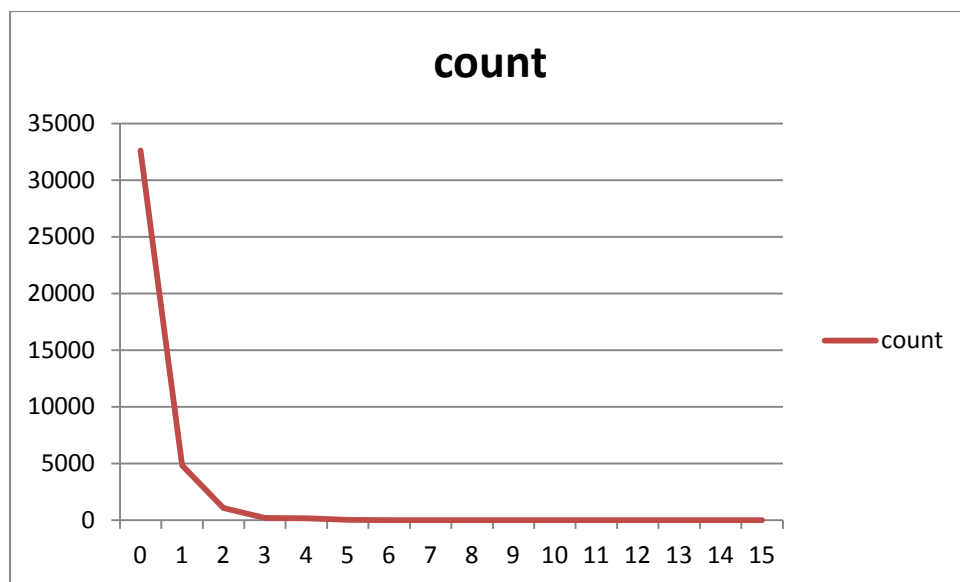


c. loadAveragePercent

loadAveragePercent	count
-1	3368964
0	32626
1	4829
2	1078
3	200
4	172

5	31
6	11
7	12
8	11
9	3
10	9
11	6
12	5
13	1
14	6
15	3

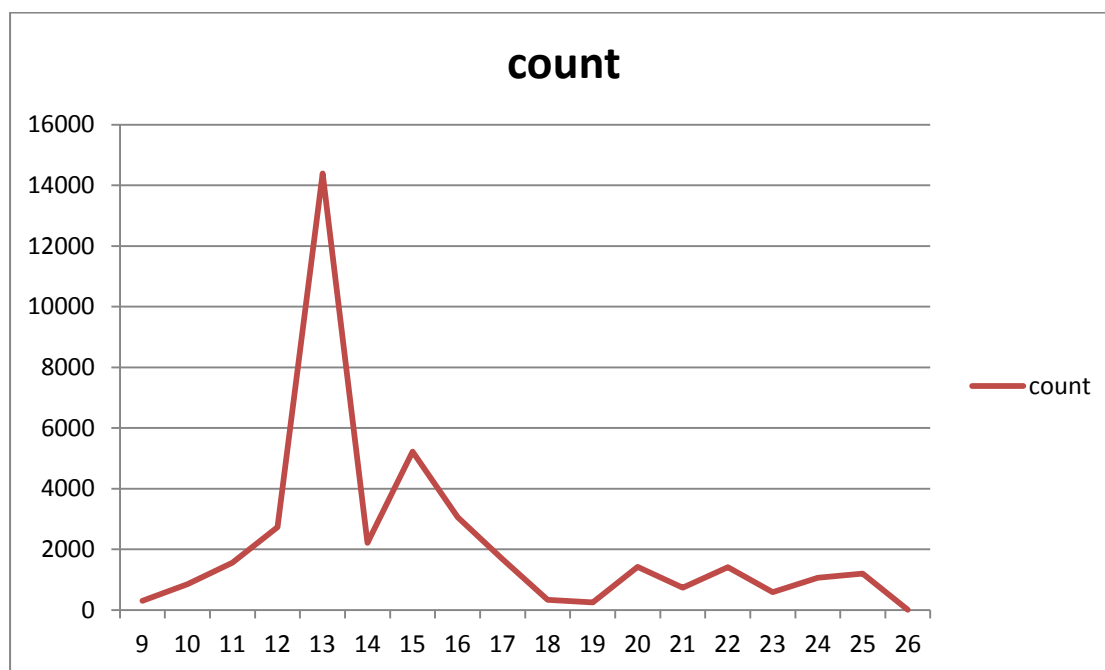
其分布如下：



d. physicalMemoryUsagePercent

physicalMemoryUsagePercent	count
-1	3368964
9	306
10	853
11	1562
12	2734
13	14389
14	2211
15	5216
16	3055
17	1673
18	334
19	253
20	1420
21	734

22	1411
23	589
24	1057
25	1202
26	4



从上面的统计结果来看，bbexport 数据中比较有分析意义的字段是 diskUsagePercent, numProcs, loadAveragePercent 和 physicalMemoryUsagePercent，其中 diskUsagePercent 的异常比较明显，出现了 90% 以上的数据；其他的字段还需要进一步的分析。

3) nf1 数据统计

由于 nf1 数据记录的是网络上数据流的信息，包含源地址，目标地址，端口号等类别型数据和发包大小这一数据型记录，所以只做了一项统计，就是一个 ip 访问过的端口号。发现有一些 ip 地址访问了上万个端口，这应该属于异常情况。这些 ip 地址如下所示：

172.30.0.3	23825
10.6.6.14	62531
10.11.6.15	62531
10.18.6.123	62531
10.16.5.15	62531
10.7.6.3	62531
10.6.6.13	62531
10.100.1.6	62531
10.10.6.2	62531
10.6.6.6	62532
10.7.7.10	62532

以上是对数据的一些统计情况。

这次的 vast challenge 3 与在阿里云进行的项目比较相似，都具有一份监控数据，其中包含了机器的一些属性，而不同之处在于，本次的数据多出了 netflow 部分，记录了网络上的包的记录。同时，数据量比较大，一周的监控数据有 340 万条，而一周的 netflow 的总量大概有 5000 万条。大的数据量也是一项挑战。

2. 开始写毕业论文。现在已经完成了前两章的内容，开始写第三章。

3. 大屏可视化的调研工作进展不大。

下周工作

1. 继续对 challenge 3 的数据进行统计分析，将所有数据导入数据库，尝试定位一些时间。

2. 写毕业论文。

3. 继续大屏可视化的调研工作。